

Asymptotic theory for maximum likelihood estimates in reduced-rank multivariate generalized linear models

E. Bura^{a,b}, S. Duarte^c, L. Forzani^c, E. Smucler^{d,e} and M. Sued^e

^aInstitute of Statistics and Mathematical Methods in Economics, TU Wien, Vienna, Austria; ^bDepartment of Statistics, George Washington University, Washington, DC, USA; ^cFacultad de Ingeniería Química, UNL, Santa Fe, Argentina; ^dDepartment of Statistics, University of British Columbia, Vancouver, BC, Canada; ^eInstituto de Cálculo, UBA, Buenos Aires, Argentina

ABSTRACT

Reduced-rank regression is a dimensionality reduction method with many applications. The asymptotic theory for reduced rank estimators of parameter matrices in multivariate linear models has been studied extensively. In contrast, few theoretical results are available for reduced-rank multivariate generalized linear models. We develop M-estimation theory for concave criterion functions that are maximized over parameter spaces that are neither convex nor closed. These results are used to derive the consistency and asymptotic distribution of maximum likelihood estimators in reduced-rank multivariate generalized linear models, when the response and predictor vectors have a joint distribution. We illustrate our results in a real data classification problem with binary covariates.

ARTICLE HISTORY

Received 26 June 2017
Accepted 29 March 2018

KEYWORDS

M-estimation; exponential family; rank restriction; non-convex; parameter spaces

1. Introduction



The multivariate multiple linear regression model for a q -dimensional response $Y = (Y_1, \dots, Y_q)^T$ and a p -dimensional predictor vector $X = (X_1, \dots, X_p)^T$ postulates that $Y = BX + \epsilon$, where B is a $q \times p$ matrix and $\epsilon = (\epsilon_1, \dots, \epsilon_q)^T$ is the error term, with $E(\epsilon) = 0$ and $\text{var}(\epsilon) = \Sigma$. Based on a random sample $(X_1, Y_1), \dots, (X_n, Y_n)$ satisfying the model, ordinary least squares estimates the parameter matrix B by minimizing the squared error loss function $\sum_{i=1}^n \|Y_i - BX_i\|^2$ to obtain


$$\hat{B}_{\text{OLS}} = \left(\sum_{i=1}^n Y_i X_i^T \right) \left(\sum_{i=1}^n X_i X_i^T \right)^{-1}. \quad (1)$$

Reduced-rank regression introduces a rank constraint on B , so that Equation (1) is minimized subject to the constraint $\text{rank}(B) \leq r$, where $r < \min(p, q)$. The solution is $\hat{B}_{\text{RRR}}^T = \hat{B}_{\text{OLS}}^T U_r U_r^T$, where U_r are the first r singular vectors of $\hat{Y}^T = \hat{B}_{\text{OLS}} X^T$ (see, e.g. [1]).

Reduced-rank regression has attracted attention as a regularisation method by introducing a shrinkage penalty on B . Moreover, it is used as a dimensionality reduction method as it constructs latent factors in the predictor space that explain the variance of the responses.

Anderson [2] obtained the likelihood-ratio test of the hypothesis that the rank of B is a given number and derived the associated asymptotic theory under the assumption of normality of Y and

CONTACT E. Bura  efstathia.bura@tuwien.ac.at  Institute of Statistics and Mathematical Methods in Economics, TU Wien, A-1040 Vienna, Austria; Department of Statistics, George Washington University, Washington, DC 20052, USA

 Supplemental data for this article can be accessed at <https://doi.org/10.1080/02331888.2018.1467420>

non-stochastic X . Under the assumption of joint normality of Y and X , Izenman [3] obtained the asymptotic distribution of the estimated reduced-rank regression coefficient matrix and drew connections between reduced-rank regression, principal component analysis and correlation analysis. Specifically, principal component analysis coincides with reduced-rank regression when $Y = X$ [4]. Recently, Fan et al. [5] studied the theoretical properties of nuclear norm regularized maximum likelihood type estimates in a class of models that includes reduced-rank regression. They derive statistical rates of convergence in possibly high-dimensional scenarios. However, they do not provide asymptotic results that can be used to construct confidence intervals or hypothesis tests. The monograph by Reinsel and Velu [1] contains a comprehensive survey of the theory and history of reduced rank regression, including in time series, and its many applications.

Despite the application potential of reduced-rank regression, it has received limited attention in generalized linear models. Yee and Hastie [6] were the first to introduce reduced-rank regression to the class of multivariate generalized linear models, which covers a wide range of data types for both the response and predictor vectors, including categorical data. Multivariate or vector generalized linear models (VGLMs) is the topic of Yee's [7] book, which is accompanied by the associated R packages, `VGAM` and `VGAMdata`. Yee and Hastie [6] proposed an alternating estimation algorithm, which was shown to result in the maximum likelihood estimate of the parameter matrix in reduced-rank multivariate generalized linear models by Bura et al. [8]. Asymptotic theory for the restricted rank maximum likelihood estimates of the parameter matrix in multivariate GLMs has not been developed yet.

In general, a maximum likelihood estimator is a concave M-estimator in the sense that it maximizes the empirical mean of a concave criterion function. Asymptotic theory for M-estimators defined through a concave function has received much attention. Huber [9], Haberman [10] and Niemiro [11] are among the classical references. More recently, Hjort and Pollard [12] presented a unified framework for the statistical theory of M-estimation for convex criterion functions that are minimized over open convex sets of a Euclidean space. Geyer [13] studied M-estimators restricted to a closed subset of a Euclidean space.

The rank restriction in reduced rank regression imposes constraints that have not been studied before in M-estimation as they result in neither convex nor closed parameter spaces. In this paper we (a) develop M-estimation theory for concave criterion functions, which are maximized over parameter spaces that are neither convex nor closed, and (b) apply the results from (a) to obtain asymptotic theory for reduced rank regression estimators in generalized linear models. Specifically, we derive the asymptotic distribution and properties of maximum likelihood estimators in reduced-rank multivariate generalized linear models where both the response and predictor vectors have a joint distribution. The asymptotic theory we develop covers reduced-rank regression for linear models as a special case. We show the improvement in inference the asymptotic theory offers via analysing the data set Yee and Hastie [6] analysed.

Throughout, for a function $f : \mathbb{R}^q \rightarrow \mathbb{R}$, $\nabla f(x)$ denotes the row vector $\nabla f(x) = (\partial f(x)/\partial x_1, \dots, \partial f(x)/\partial x_q)$, $\dot{f}(x)$ stands for the column vector of derivatives, while $\nabla^2 f(x)$ denotes the symmetric matrix of second-order derivatives. For a vector valued function $f : \mathbb{R}^{q_1} \rightarrow \mathbb{R}^{q_2}$, ∇f denotes the $q_2 \times q_1$ matrix,

$$\nabla f(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \dots & \frac{\partial f_1}{\partial x_{q_1}} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \dots & \frac{\partial f_2}{\partial x_{q_1}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial f_{q_2}}{\partial x_1} & \frac{\partial f_{q_2}}{\partial x_2} & \dots & \frac{\partial f_{q_2}}{\partial x_{q_1}} \end{pmatrix}.$$

2. M-estimators

Let Z be a random vector taking values in a measurable space \mathcal{Z} and distributed according to the law P . We are interested in estimating a finite dimensional parameter $\xi_0 = \xi_0(P)$ using n independent and identically distributed copies Z_1, Z_2, \dots, Z_n of Z . In the sequel, we use Pf to denote the mean of $f(Z)$; i.e. $Pf = E[f(Z)]$.

Let Ξ be a subset of a Euclidean space and $m_\Xi : \mathcal{Z} \rightarrow \mathbb{R}$ be a known function. Assume that the parameter of interest ξ_0 is the maximizer of the map $\xi \mapsto Pm_\xi$ defined on Ξ . One can estimate ξ_0 by maximizing an empirical version of the optimization criterion. Specifically, given a known function $m_\xi : \mathcal{Z} \mapsto \mathbb{R}$, define

$$M(\xi) := Pm_\xi \quad \text{and} \quad M_n(\xi) := P_n m_\xi, \tag{2}$$

where, here and throughout, P_n denotes the empirical mean operator $P_n V = n^{-1} \sum_{i=1}^n V_i$. Hereafter, $M_n(\xi)$ and $P_n m_\xi$ will be used interchangeably as the criterion function, depending on which of the two is appropriate in a given setting.

Assume that ξ_0 is the unique maximizer of the deterministic function M defined in Equation (2). An *M-estimator* for the criterion function M_n over Ξ is defined as

$$\hat{\xi}_n = \hat{\xi}_n(Z_1, \dots, Z_n) \quad \text{maximizing} \quad M_n(\xi) = \frac{1}{n} \sum_{i=1}^n m_\xi(Z_i) \quad \text{over} \quad \Xi. \tag{3}$$

If the maximum of the criterion function M_n over Ξ is not attained but the supremum of M_n over Ξ is finite, any value $\hat{\xi}_n$ that almost maximizes the criterion function, in the sense that it satisfies

$$M_n(\hat{\xi}_n) \geq \sup_{\xi \in \Xi} M_n(\xi) - A_n, \tag{4}$$

for A_n small, can be used instead.

Definition 2.1: An estimator $\hat{\xi}_n$ that satisfies Equation (4) with $A_n = o_p(1)$ is called a weak M-estimator for the criterion function M_n over Ξ . When $A_n = o_p(n^{-1})$, $\hat{\xi}_n$ is called a strong M-estimator.

Proposition A.1 in the [Appendix](#) lists the conditions for the existence, uniqueness and strong consistency of an M-estimator, as defined in Equation (3), when M_n is concave and the parameter space is convex. Under regularity conditions, as those stated in Theorem 5.23 in van der Vaart [14], the asymptotic expansion and distribution of a consistent strong M-estimator [see Definition 2.1] for ξ_0 is given by

$$\sqrt{n}(\hat{\xi}_n - \xi_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n IF_{\xi_0}(Z_i) + o_p(1), \tag{5}$$

where $IF_{\xi_0}(Z_i) = -V_{\xi_0}^{-1} \dot{m}_{\xi_0}(Z_i)$, and V_{ξ_0} is the non-singular symmetric second derivative matrix of $M(\xi)$ at ξ_0 .

2.1. Restricted M-estimators

We now consider the optimization of M_n over $\Xi^{\text{res}} \subset \Xi$, where Ξ^{res} is the image of a function that is not necessarily injective. Specifically, we restrict the optimization problem to the set Ξ^{res} by requiring:

Condition 2.2: *There exists an open set $\Theta \subset \mathbb{R}^q$ and a map $g : \Theta \rightarrow \Xi$ such that $\xi_0 \in g(\Theta) = \Xi^{\text{res}}$.*

Even when an M-estimator for the unrestricted problem as defined in (3) exists, there is no a priori guarantee that the supremum is attained when considering the restricted problem. Nevertheless, Lemma 2.3 establishes the existence of a restricted strong M-estimator, regardless of

whether the original M-estimator is a real maximizer, weak or strong. All proofs are provided in the [Appendix](#).

Lemma 2.3: *Assume there exists a weak/strong M-estimator $\hat{\xi}_n$ for the criterion function M_n over Ξ . If Condition 2.2 holds, then there exists a strong M-estimator $\hat{\xi}_n^{\text{res}}$ for the criterion function M_n over Ξ^{res} .*

Proposition 2.4 next establishes the existence and consistency of a weak restricted M-estimator sequence when $m_\xi(z)$ is a concave function in ξ under the same assumptions as those of the unrestricted problem, stated in Proposition A.1 in the [Appendix](#).

Proposition 2.4: *Assume that Condition 2.2 holds. Then, under the assumptions of Proposition A.1 in the [Appendix](#), there exists a strong M-estimator of the restricted problem over Ξ^{res} . Moreover, any strong M-estimator of the restricted problem converges to ξ_0 in probability.*

We derive next the asymptotic distribution of $\hat{\xi}_n^{\text{res}} = g(\hat{\theta}_n)$, with $\hat{\theta}_n \in \Theta$. The constrained estimator $\hat{\xi}_n^{\text{res}}$ is well defined under Lemma 2.3, even when $\hat{\theta}_n$ is not uniquely determined. If $\hat{\theta}_n$ were unique and $\sqrt{n}(\hat{\theta}_n - \theta)$ had an asymptotic distribution, one could use a Taylor series expansion for g to derive asymptotic results for $\hat{\xi}_n^{\text{res}}$. Building on this idea, Condition 2.5 introduces a parametrization of a neighbourhood of ξ_0 that allows applying standard tools in order to obtain the asymptotic distribution of the restricted M-estimator.

Condition 2.5: *Given $\xi_0 \in g(\Theta)$, there exists an open set \mathcal{M} in Ξ^{res} with $\xi_0 \in \mathcal{M} \subset g(\Theta)$, and (\mathcal{S}, h) , where \mathcal{S} is an open set in \mathbb{R}^{q_s} , $q_s \leq q$, and $h : \mathcal{S} \rightarrow \mathcal{M}$ is one-to-one, bi-continuous and twice continuously differentiable, with $\xi_0 = h(s_0)$ for some $s_0 \in \mathcal{S}$.*

Under the setting in Condition 2.5, we will prove that $\hat{s}_n = h^{-1}(\hat{\xi}_n^{\text{res}})$ is a strong M-estimator for the criterion function $P_n m_{h(s)}$ over \mathcal{S} . Then, we can apply Theorem 5.23 of van der Vaart [14] to obtain the asymptotic behaviour of \hat{s}_n , which, combined with a Taylor expansion of h about s_0 , yield a linear expansion for $\hat{\xi}_n^{\text{res}}$. Finally, requiring Condition 2.6, which relates the parametrizations (\mathcal{S}, h) and (Θ, g) , suffices to derive the asymptotic distribution of $\hat{\xi}_n^{\text{res}}$ in terms of g .

Condition 2.6: *Consider (Θ, g) as in Condition 2.2 and (\mathcal{S}, h) as in Condition 2.5. For each $\theta_0 \in g^{-1}(\xi_0)$, $\text{span} \nabla g(\theta_0) = \text{span} \nabla h(s_0)$.*

Condition 2.6 ensures that $T_{\xi_0} = \text{span} \nabla g(\theta_0)$ is well defined regardless of the fact that $g^{-1}(\xi_0)$ may contain multiple θ_0 's. Moreover, T_{ξ_0} also agrees with $\text{span} \nabla h(s_0)$. Consequently, the orthogonal projection $\Pi_{\xi_0(\Sigma)}$ onto T_{ξ_0} with respect to the inner product defined by a symmetric positive definite matrix Σ satisfies

$$\begin{aligned} \Pi_{\xi_0(\Sigma)} &= \nabla g(\theta_0)(\nabla g(\theta_0)^T \Sigma \nabla g(\theta_0))^\dagger \nabla g(\theta_0)^T \Sigma \\ &= \nabla h(s_0)(\nabla h(s_0)^T \Sigma \nabla h(s_0))^{-1} \nabla h(s_0)^T \Sigma, \end{aligned} \tag{6}$$

where A^\dagger denotes a generalized inverse of the matrix A . The gradient of g is not necessarily of full rank, in contrast to the gradient of h . Note that $\Pi_{\xi_0(\Sigma)}$ is idempotent ($\Pi_{\xi_0(\Sigma)}^2 = \Pi_{\xi_0(\Sigma)}$) and the span of its columns is equal to T_{ξ_0} . However, $\Pi_{\xi_0(\Sigma)}$ is not in general symmetric, but rather self-adjoint with respect to the inner product induced by Σ since $\langle x, y \rangle_\Sigma := y^T \Sigma x$.

Proposition 2.7: Assume that Conditions 2.2, 2.5 and 2.6 hold. Assume also that the unrestricted problem satisfies the regularity Conditions A.2–A.4 in the Appendix. Then, any strong M-estimator $\hat{\xi}_n^{\text{res}}$ of the restricted problem that converges in probability to ξ_0 satisfies

$$\sqrt{n}(\hat{\xi}_n^{\text{res}} - \xi_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \Pi_{\xi_0(-V_{\xi_0})} IF_{\xi_0}(Z_i) + o_p(1), \tag{7}$$

where $IF_{\xi_0}(Z_i) = -V_{\xi_0}^{-1} \dot{m}_{\xi_0}(Z_i)$ is the influence function of the unrestricted estimator defined in Equation (5), V_{ξ_0} is the non-singular symmetric second derivative matrix of $M(\xi)$ at ξ_0 , and $\Pi_{\xi_0(-V_{\xi_0})}$ is defined according to Equation (6).

Moreover, $\sqrt{n}(\hat{\xi}_n^{\text{res}} - \xi_0)$ is asymptotically normal with mean zero and asymptotic variance

$$\text{avar}\{\sqrt{n}(\hat{\xi}_n^{\text{res}} - \xi_0)\} = \Pi_{\xi_0(-V_{\xi_0})} V_{\xi_0}^{-1} P \left\{ \dot{m}_{\xi_0} \dot{m}_{\xi_0}^T \right\} V_{\xi_0}^{-1} \Pi_{\xi_0(-V_{\xi_0})}^T. \tag{8}$$

As an aside remark, we conjecture that for estimators that are maximizers of a criterion function under restrictions that satisfy Conditions 2.2–2.6, when Equation (5) is true, Equation (7) also holds. This can be important since the asymptotic distribution of restricted estimators will be derived directly from the asymptotic distribution of the unrestricted one.

The optimization problem that defines the restricted M-estimators considered in this paper is, in general, not convex and hence difficult to solve. However, in the case of maximum likelihood estimation in reduced rank multivariate generalized linear models, the main application of the results in our paper, efficient algorithms exist for solving it (see [6] and the VGAM R package).

3. Asymptotic theory for the maximum likelihood estimator in reduced rank multivariate generalized linear models

In this section we show that maximum likelihood estimators in reduced-rank multivariate generalized linear models are restricted strong M-estimators for the conditional log-likelihood. Using results in Section 2.1, we obtain the existence, consistency and asymptotic distribution of maximum likelihood estimators in reduced-rank multivariate generalized linear models. In practice, these estimators can be obtained using the R package VGAM, developed by Yee [15].

3.1. Exponential family

Let $Y = (Y_1, \dots, Y_q)^T$ be a q -dimensional random vector and assume that its distribution belongs to a k -parameter canonical exponential family with pdf (pms)

$$f_{\eta}(y) = \exp\{\eta^T T(y) - \psi(\eta)\} h(y), \tag{9}$$

where $T(y) = (T_1(y), \dots, T_k(y))^T$ is a vector of known real-valued functions, $h(y) \geq 0$ is a non-negative known function and $\eta \in \mathbb{R}^k$ is the vector of natural parameters, taking values in

$$H = \{\eta \in \mathbb{R}^k : \int \exp\{\eta^T T(y)\} h(y) dy < \infty\}, \tag{10}$$

where the integral is replaced by a sum when Y is discrete. The set H of the natural parameter space is assumed to be open and convex in \mathbb{R}^k , and ψ a strictly convex function defined on H . Moreover, we assume $\psi(\eta)$ is convex and infinitely differentiable in H . In particular,

$$\nabla \psi(\eta) = E_{\eta}^T(T(Y)) \quad \text{and} \quad \nabla^2 \psi(\eta) = \text{var}_{\eta}(T(Y)), \quad \text{for every } \eta \in H, \tag{11}$$

where $\nabla^2 \psi$ is the $k \times k$ matrix of second derivatives of ψ . Since ψ is strictly convex, $\text{var}_{\eta}(T(Y))$ is non-singular for every $\eta \in H$.

3.2. Multivariate generalized linear models

Let $Z = (X, Y)$ be a random vector, where now $Y \in \mathbb{R}^q$ is a multivariate response and $X \in \mathbb{R}^p$ is a vector of predictors. The multivariate generalized linear model postulates that the conditional distribution of Y given X belongs to some fixed exponential family and hypothesizes that the k -vector of natural parameters, which we henceforth call η_x to emphasize the dependence on x , depends linearly on the vector of predictors. Thus, the pdf (pms) of $Y \mid X = x$ is given by

$$f_{Y|X=x}(y) = \exp\{\eta_x^T T(y) - \psi(\eta_x)\}h(y), \tag{12}$$

where $\eta_x \in \mathbb{R}^k$ depends linearly on x .

Frequently, a subset of the natural parameters depends on x , whereas its complement does not. The normal linear model with constant variance is such an example. To accommodate this structure, we partition the vector η_x indexing model (12) into η_{x1} and η_{x2} , with k_1 and k_2 components, and assume that H , the natural parameter space of the exponential family, is $\mathbb{R}^{k_1} \times H_2$, where H_2 is an open convex subset of \mathbb{R}^{k_2} . We also assume that

$$\eta_x = \begin{pmatrix} \eta_{x1} \\ \eta_{x2} \end{pmatrix} = \begin{pmatrix} \bar{\eta}_1 + \beta x \\ \bar{\eta}_2 \end{pmatrix}, \tag{13}$$

where $\beta \in \mathbb{R}^{k_1 \times p}$, $\bar{\eta}_1 \in \mathbb{R}^{k_1}$ and $\bar{\eta}_2 \in H_2$. Let $\xi = (\bar{\eta}_1^T, \text{vec}^T(\beta), \bar{\eta}_2^T)^T \in \Xi = \mathbb{R}^{k_1} \times \mathbb{R}^{k_1 p} \times H_2$, denote a generic vector and ξ_0 the true parameter. Suppose n independent and identically distributed copies of $Z = (X, Y)$ satisfying Equations (12) and (13), with true parameter vector ξ_0 , are available. Given a realization $z_i = (x_i, y_i)$, $i = 1, \dots, n$, the conditional log-likelihood, up to a factor that does not depend on the parameter of interest, is

$$\mathcal{L}_n(\beta; \bar{\eta}) = \frac{1}{n} \sum_{i=1}^n \left\{ \eta_{x_i}^T T(y_i) - \psi(\eta_{x_i}) \right\}. \tag{14}$$

Let

$$m_{(\beta; \bar{\eta})}(z) = \eta_x^T T(y) - \psi(\eta_x). \tag{15}$$

By definition (3), the maximum likelihood estimator (MLE) of the parameter indexing model (12) subject to (13), if it exists, is an M-estimator.

Theorem 3.1 next establishes the existence, consistency and asymptotic normality of $\hat{\xi}_n$, the MLE of ξ_0 .

Theorem 3.1: *Assume that $Z = (X, Y)$ satisfies model (12) subject to (13) with true parameter ξ_0 . Under regularity conditions (A20), (A21), (A22) and (A23) in the Appendix, the maximum likelihood estimate of ξ_0 , $\hat{\xi}_n$, exists, is unique and converges in probability to ξ_0 . Moreover, $\sqrt{n}(\hat{\xi}_n - \xi_0)$ is asymptotically normal with covariance matrix*

$$W_{\xi_0} = \left[E \left\{ F(X)^T \nabla^2 \psi(F(X)\xi_0) F(X) \right\} \right]^{-1}, \tag{16}$$

where

$$F(x) = \begin{pmatrix} ((1, x^T) \otimes I_{k_1}) & 0 \\ 0 & I_{k_2} \end{pmatrix},$$

and $\nabla^2 \psi$ was defined in Equation (11).

3.3. Partial reduced rank multivariate generalized linear models

When the number of natural parameters or the number of predictors is large, the precision of the estimation and/or the interpretation of results can be adversely affected. A way to address this is to assume that the parameters live in a lower dimensional space. That is, we assume that the vector of predictors can be partitioned as $x = (x_1^T, x_2^T)^T$ with $x_1 \in \mathbb{R}^r$ and $x_2 \in \mathbb{R}^{p-r}$, and that the parameter corresponding to x_1 , $\beta_1 \in \mathbb{R}^{k_1 \times r}$, has rank $d < \min\{k_1, r\}$. In this way, the natural parameters η_x in Equation (12) are related to the predictors via

$$\eta_x = \begin{pmatrix} \eta_{x1} \\ \eta_{x2} \end{pmatrix} = \begin{pmatrix} \bar{\eta}_1 + \beta_1 x_1 + \beta_2 x_2 \\ \bar{\eta}_2 \end{pmatrix}, \tag{17}$$

where $\beta_1 \in \mathbb{R}_d^{k_1 \times r}$, the set of matrices in $\mathbb{R}^{k_1 \times r}$ of rank $d \leq \min\{k_1, r\}$, while $\beta_2 \in \mathbb{R}^{k_1 \times (p-r)}$, and $\beta = (\beta_1, \beta_2)$.

Following Yee and Hastie [6], we refer to the exponential conditional model (12) subject to the restrictions imposed in Equation (17) as *partial reduced rank multivariate generalized linear model*. The reduced-rank multivariate generalized linear model is a special case of model (17) with $\beta_2 = 0$.

To obtain the asymptotic distribution of the M-estimators for this reduced model, we will show that Conditions 2.2–2.6 are satisfied for Ξ^{res} , (Θ, g) , \mathcal{M} and (\mathcal{S}, h) , which are defined next. To maintain consistency with notation introduced in Section 2.1, we vectorize each matrix involved in the parametrization of our model and reformulate the parameter space accordingly for each vectorized object. With this understanding, we use the symbol \cong to indicate that a matrix space component in a product space is identified with its image through the operator $\text{vec} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{mn}$. In the sequel, to keep the notation as simple as possible, we concatenate column vectors without transposing them; that is, we write (a, b) for $(a^T, b^T)^T$. Moreover, we write $\xi = (\bar{\eta}_1, \beta, \bar{\eta}_2)$, with the understanding that β stands for $\text{vec}(\beta)$.

For the non-restricted problem, β_1 belongs to $\mathbb{R}^{k_1 \times r}$, so that the entire parameter $\xi = (\bar{\eta}_1, \beta_1, \beta_2, \bar{\eta}_2)$ belongs to

$$\Xi \cong \mathbb{R}^{k_1} \times \mathbb{R}^{k_1 \times r} \times \mathbb{R}^{k_1 \times (p-r)} \times H_2. \tag{18}$$

However, for the restricted problem, we assume that the true parameter $\xi_0 = (\bar{\eta}_{01}, \beta_{01}, \beta_{02}, \bar{\eta}_{02})$ belongs to

$$\Xi^{\text{res}} \cong \mathbb{R}^{k_1} \times \mathbb{R}_d^{k_1 \times r} \times \mathbb{R}^{k_1 \times (p-r)} \times H_2. \tag{19}$$

Let

$$\Theta \cong \mathbb{R}^{k_1} \times \left\{ \mathbb{R}_d^{k_1 \times d} \times \mathbb{R}_d^{d \times r} \right\} \times \mathbb{R}^{k_1 \times (p-r)} \times H_2 \tag{20}$$

and consider $g : \Theta \rightarrow \Xi^{\text{res}}$, with $(\bar{\eta}_1, (S, T), \beta_2, \bar{\eta}_2) \mapsto (\bar{\eta}_1, ST, \beta_2, \bar{\eta}_2)$. Without loss of generality, we assume that $\beta_{01} \in \mathbb{R}_{\text{first},d}^{k_1 \times r}$, the set of matrices in $\mathbb{R}_d^{k_1 \times r}$ whose first d rows are linearly independent. Therefore,

$$\beta_{01} = \begin{pmatrix} I_d \\ A_0 \end{pmatrix} B_0, \quad \text{with } A_0 \in \mathbb{R}^{(k_1-d) \times d} \quad \text{and} \quad B_0 \in \mathbb{R}_d^{d \times r}. \tag{21}$$

This is trivial since $\beta_{01} \in \mathbb{R}_d^{k_1 \times r}$ and its first d rows are linearly independent. Consider

$$\mathcal{M} \cong \mathbb{R}^{k_1} \times \mathbb{R}_{\text{first},d}^{k_1 \times r} \times \mathbb{R}^{k_1 \times (p-r)} \times H_2, \quad \text{and so} \quad \xi_0 \in \mathcal{M} \subset \Xi^{\text{res}}. \tag{22}$$

Finally, let

$$\mathcal{S} \cong \mathbb{R}^{k_1} \times \left\{ \mathbb{R}^{(k_1-d) \times d} \times \mathbb{R}_d^{d \times r} \right\} \times \mathbb{R}^{k_1 \times (p-r)} \times H_2, \tag{23}$$

and $h : S \rightarrow \mathcal{M}$ be the map

$$(\bar{\eta}_1, (A, B), \beta_2, \bar{\eta}_2) \mapsto \left(\bar{\eta}_1, \begin{pmatrix} B \\ AB \end{pmatrix}, \beta_2, \bar{\eta}_2 \right). \tag{24}$$

Proposition 3.2: *Conditions 2.2–2.6 are satisfied for $\xi_0, \Xi, \Xi^{\text{res}}, (\Theta, g), \mathcal{M}$ and (S, h) defined in Equations (18)–(24), respectively.*

Under the rank restriction on β_1 in Equation (17), the existence of the maximum likelihood estimator cannot be guaranteed in the sense of an M-estimator as defined in Equation (3) with $m_\xi = m_{(\beta; \bar{\eta})}$ in Equation (15), and Ξ replaced by Ξ^{res} in Equation (18). However, we can work with a strong M-estimator sequence for the criterion function $P_n m_{(\beta; \bar{\eta})}$ over Ξ^{res} using Lemma 2.3. Theorem 3.3 states our main contribution.

Theorem 3.3: *Let $\xi_0 = (\bar{\eta}_{01}, \beta_{01}, \beta_{02}, \bar{\eta}_{02})$ denote the true parameter value of $\xi = (\bar{\eta}_1, \beta_1, \beta_2, \bar{\eta}_2)$. Assume that $Z = (X, Y)$ satisfies model (12) subject to (17), with $\xi_0 \in \Xi^{\text{res}}$ defined in Equation (19). Then, there exists a strong maximizing sequence for the criterion function $P_n m_\xi$ over Ξ^{res} for $m_\xi = m_{(\beta; \bar{\eta})}$ defined in Equation (15). Moreover, any weak M-estimator sequence $\{\hat{\xi}_n^{\text{res}}\}$ converges to ξ_0 in probability.*

If $\{\hat{\xi}_n^{\text{res}}\}$ is a strong M-estimator sequence, then $\sqrt{n}(\hat{\xi}_n^{\text{res}} - \xi_0)$ is asymptotically normal with covariance matrix

$$\begin{aligned} \text{avar}\{\sqrt{n}(\hat{\xi}_n^{\text{res}} - \xi_0)\} &= \Pi_{\xi_0(W_{\xi_0}^{-1})} W_{\xi_0} \\ &= G(G^T W_{\xi_0}^{-1} G)^{-1} G^T, \end{aligned} \tag{25}$$

where W_{ξ_0} is defined in Equation (16) and

$$G = \begin{pmatrix} I_{k_1} & 0 & 0 & 0 & 0 \\ 0 & B_0^T \otimes I_{k_1} & I_r \otimes C_0 & 0 & 0 \\ 0 & 0 & 0 & I_{k_1(p-r)} & 0 \\ 0 & 0 & 0 & 0 & I_{k_2} \end{pmatrix}, \tag{26}$$

with $\beta_{01} = C_0 B_0, C_0 \in \mathbb{R}^{k_1 \times d}$ and $B_0 \in \mathbb{R}^{d \times r}$ any decomposition of β_{01} .

Remark 3.4: The asymptotic variance of the estimator does not depend on the specific decomposition of β_{01} . That is, for another $\beta_{01} = \tilde{C}_0 \tilde{B}_0$, even if (26) changes, (25) remains the same.

Remark 3.5: A plug-in procedure can be used to estimate the elements of the matrix in Equation (25) and then, appealing to Slutsky’s Theorem, asymptotic confidence regions for $\hat{\xi}_n^{\text{res}}$ can be constructed. Also, bootstrap variance estimates can be computed by sampling with replacement from the empirical distribution, leading to so-called normal bootstrap intervals (see, e.g. [16], Section 8.3).

Remark 3.6: Since $\Pi_{\xi_0(W_{\xi_0}^{-1})}$ is a projection, $\Pi_{\xi_0(W_{\xi_0}^{-1})} W_{\xi_0} \leq W_{\xi_0}$. That is, the eigenvalues of $W_{\xi_0} - \Pi_{\xi_0(W_{\xi_0}^{-1})} W_{\xi_0}$ are non-negative, so that using partial reduced-rank multivariate generalized linear models results in efficiency gain.

Remark 3.7: Decomposing x into more sets of predictors with reduced rank parameters amounts to adding new rows, like the second set of rows of G in Equation (26), for each reduced rank parameter.

Remark 3.8: For more general families, like the one considered by Yee and Hastie [6], if the family in question satisfies the regularity conditions in Proposition 2.7, the asymptotic variance of the estimators can also be computed by making the corresponding substitutions in the formulas of Proposition 2.7.

4. Application: marital status in a workforce study

Yee and Hastie [6] analyse data from a self-administered questionnaire collected in a large New Zealand workforce observational study conducted during 1992–1993. For homogeneity, the analysis was restricted to a subset of 4105 European males with no missing values in any of the variables used. Yee and Hastie [6] were interested in exploring whether certain lifestyle and psychological variables were associated with marital status, especially separation/divorce. The response variable is $Y = \text{maritalstatus}$, with levels 1 = single, 2 = separatedordivorced, 3 = widower, and 4 = marriedorlivingwithapartner. The married/partnered are the reference group. Data on 14 regressors were collected, 12 of which are binary (1/0 for presence/absence, respectively). These have been coded so that their presence is negative healthwise. Their goal was to investigate if and how these 12 *unhealthy* variables were related to Y , adjusting for age and level of education. The variables are described in Table 1.

A categorical response Y taking values 1,2,3,4 with probabilities $\text{pr}(Y = i) = p_i$ can be expressed as a multinomial vector to fit the generalized linear model presented in this paper, $Y = (Y_1, Y_2, Y_3, Y_4)^T$, where $Y_i = 1$ if $Y = i$ and $Y_i = 0$ otherwise, and $\sum_{i=1}^4 Y_i = 1$. The pdf of Y can be written as

$$f_Y(y) = \exp\{\eta^T T(y) - \psi(\eta)\}h(y), \tag{27}$$

where $y = (y_1, y_2, y_3, y_4)$, $T(y) = (y_1, y_2, y_3)$, $\eta = (\eta_1, \eta_2, \eta_3) \in \mathbb{R}^3$, $\psi(\eta) = 1 + \sum_{i=1}^3 \exp(\eta_i)$ and $h(y) = 1/(y_1!y_2!y_3!(1 - y_1 - y_2 - y_3)!)$. The natural parameter $\eta = (\eta_1, \eta_2, \eta_3) \in \mathbb{R}^3$, is related to the pdf of Y through the identity $\eta_i = \log(p_i/p_4)$, $i = 1,2,3$.

Let X be the vector of predictor variables and consider $p_i = p_i(x) = \text{pr}(Y_i = 1|X = x)$. The dependence of p_i on x will not be made explicit in the notation. As in [6] we fit a multinomial regression model to $Y | X = x$, as in Equation (27), with

$$\eta_x = \begin{pmatrix} \log(p_1/p_4) \\ \log(p_2/p_4) \\ \log(p_3/p_4) \end{pmatrix} = \bar{\eta} + \beta x, \tag{28}$$

where $\bar{\eta} \in \mathbb{R}^3$ is the intercept and $\beta \in \mathbb{R}^{3 \times 14}$ is the coefficient matrix, so that there are $3 \times 14 + 3 = 42 + 3 = 45$ parameters to estimate. When a multinomial linear model is fitted to the data at level 0.05, `age30` and `binge` are significant for $\log(p_1/p_4)$, `smokehow` and `tense` for $\log(p_2/p_4)$, and only `age30` is significant for $\log(p_3/p_4)$.

Table 1. Variables used in the workforce study.

Variable name	Description
<code>marital</code>	Marital status. 1 = single, 2 = separated or divorced, 3 = widower, and 4 = married or living with a partner
<code>age30</code>	Age –30, in years
<code>logedul</code>	$\log(1+$ years of education at secondary or high school)
<code>binge</code>	In the last 3 months what is the largest number of drinks that you had on any one day? (1 = 20 or more, 0 = less than 20)
<code>smokenow</code>	Current smoker?
<code>sun</code>	Does not usually wear a hat, shirt or suntan lotion when outside during summer
<code>nerves</code>	Do you suffer from ‘nerves’?
<code>nervous</code>	Would you call yourself a ‘nervous’ person?
<code>hurt</code>	Are your feelings easily hurt?
<code>tense</code>	Would you call yourself tense or ‘highly strung’?
<code>miserable</code>	Do you feel ‘just miserable’ for no reason?
<code>fedup</code>	Do you often feel ‘fed-up’?
<code>worry</code>	Do you worry about awful things that might happen?
<code>worrier</code>	Are you a worrier?
<code>mood</code>	Does your mood often go up and down?

Table 2. Estimators with standard errors for β_1 for the generalized linear model (first two rows) and the reduced generalized linear model (last two rows).

Variable		$\log(p_1/p_4)$	$\log(p_2/p_4)$	$\log(p_3/p_4)$
Intercept	GLM	-1.573* (0.388)	-2.921* (0.396)	-6.123* (1.106)
	PRR-GLM	-1.762* (0.377)	-2.699* (0.383)	-6.711* (1.047)
age30	GLM	-0.190* (0.008)	0.012 (0.008)	0.077* (0.024)
	PRR-GLM	-0.191* (0.008)	0.012 (0.008)	0.086* (0.023)
logedu1	GLM	0.254 (0.228)	-0.316 (0.214)	-0.198 (0.566)
	PRR-GLM	0.338 (0.225)	-0.365 (0.213)	-0.089 (0.559)

* Significance at 5% level.

We next fitted a partial reduced rank multivariate generalized linear model, where the two continuous variables, `age30` and `logedu1`, were not subject to restriction. That is,

$$\eta_x = \bar{\eta} + \beta x = \bar{\eta} + \beta_1 x_1 + \beta_2 x_2, \tag{29}$$

where x_2 represent the continuous variables and x_1 the 12 binary predictors. The AIC criterion estimates the rank of β_1 in Equation (29) to be one (see [6]). Using the asymptotic results from the current paper, Duarte [17] developed a test based on Bura and Yang [18] that also estimates the dimension to be 1. Therefore, in our notation, $q=4$, $k = k_1 = 3$, $p = 14$, $r = 2$, $d = 1$, and $\beta_1 = AB$, $A : 3 \times 1$ y $B : 1 \times 12$, $\beta_2 : 3 \times 2$ and $\bar{\eta} : 3 \times 1$. The rank restriction results in a drastic reduction in the total number of parameters from 45 to 24.

The reduction in the estimation burden is also reflected in how tight the confidence intervals are compared with those in the unrestricted model, as can be seen in Table 3 and Table 2 in [6]. As a consequence the variables `nervous`, `hurt`, which are not significant in the unrestricted generalized linear model, are significant in the reduced (29). Furthermore, some variables, such as `binge`, `smokenow`, `nervous` and `tense`, are now significant for all responses.

All significant coefficients are positive. These correspond to the variables `binge`, `smokenow`, `nervous`, `tense` and `hurt` for single and divorced/separated groups. Since the positive value of the binary variables indicates poor lifestyle and negative psychological characteristics, our analysis concludes that for men with these features, the chance of being single, divorced or widowed is higher than the chance of being married, adjusting for age and education. Also, the coefficients corresponding to the response $\log(p_3/p_4)$ are twice as large as those of $\log(p_1/p_4)$, suggesting the effect of the predictors differs in each group. All computations were performed using the R package VGAM, developed by Yee [15]. The R script to reproduce the data analysis in Section 4 can be accessed at supplemental data.

5. Discussion

With the exception of the work of Yee and his collaborators on VGLMs [6,7,19], where the distribution of the response can be any member of the multivariate exponential family, reduced-rank regression has been almost exclusively restricted to regressions with continuous response variables. Estimation methods and the corresponding software for general partial reduced rank multivariate generalized linear models were developed by Yee and Hastie [6] and Yee [7]. Yet, the distribution and statistical properties of the estimators were not obtained. In this paper we fill this gap by developing asymptotic

Table 3. MLEs with their standard errors in parentheses for the full rank generalized linear model (first two rows) and the partial reduced rank generalized linear model (last two rows).

Variable		$\log(p_1/p_4)$	$\log(p_2/p_4)$	$\log(p_3/p_4)$
binge	GLM	0.801* (0.143)	0.318 (0.256)	1.127 (0.670)
	PRR-GLM	0.569* (0.125)	0.786* (0.196)	1.114* (0.409)
smokenow	GLM	0.022 (0.126)	0.501* (0.157)	0.654 (0.469)
	PRR-GLM	0.222* (0.088)	0.306* (0.119)	0.434* (0.208)
sun	GLM	-0.066 (0.122)	0.120 (0.161)	-0.088 (0.518)
	PRR-GLM	0.011 (0.084)	0.015 (0.116)	0.021 (0.164)
nerves	GLM	-0.102 (0.138)	0.123 (0.198)	-1.456 (0.841)
	PRR-GLM	-0.054 (0.101)	-0.074 (0.139)	-0.105 (0.197)
nervous	GLM	0.297 (0.169)	0.353 (0.228)	1.007 (0.665)
	PRR-GLM	0.312* (0.124)	0.430* (0.168)	0.609* (0.291)
hurt	GLM	0.184 (0.126)	0.210 (0.167)	0.483 (0.501)
	PRR-GLM	0.180* (0.089)	0.248* (0.122)	0.352 (0.199)
tense	GLM	0.166 (0.176)	0.483* (0.214)	1.108 (0.612)
	PRR-GLM	0.302* (0.122)	0.416* (0.163)	0.590* (0.284)
miserable	GLM	-0.050 (0.138)	0.128 (0.178)	-0.093 (0.613)
	PRR-GLM	0.019 (0.094)	0.0268 (0.129)	0.038 (0.185)
fedup	GLM	0.112 (0.122)	0.249 (0.171)	-0.214 (0.548)
	PRR-GLM	0.117 (0.094)	0.161 (0.129)	0.229 (0.185)
worry	GLM	0.113 (0.145)	-0.102 (0.209)	-0.548 (0.818)
	PRR-GLM	0.003 (0.106)	0.004 (0.146)	0.005 (0.207)
worrier	GLM	-0.027 (0.131)	-0.243 (0.180)	-0.548 (0.550)
	PRR-GLM	-0.116 (0.092)	-0.160 (0.128)	-0.227 (0.193)
mood	GLM	-0.111 (0.123)	0.092 (0.171)	-0.193 (0.553)
	PRR-GLM	-0.037 (0.087)	-0.052 (0.120)	-0.073 (0.172)

* Significant at 5% level.

theory for the restricted rank maximum likelihood estimates of the parameter matrix in multivariate GLMs.

To illustrate the potential impact of our results, we refer to the real data analysis example in Section 4. In order to assess the significance of the predictors, Yee and Hastie [6] calculate the standard errors for the coefficient matrix factors, A and B , independently and can only infer about the significance of the components of the matrix A and the components of the matrix B separately. The asymptotic distribution for either estimator is obtained assuming that the other is fixed and known. In this way, they first analyse $v = Bx_1$ to check which predictors are significant and then Av to examine how they influence each response. Their standard errors are ad-hoc and it is unclear what the product of standard errors measures as relates to the significance of the product of the components of the coefficient matrix $\beta_1 = AB$. Moreover, this practical ad-hoc approach cannot readily be extended when $d > 1$.

Using the results of Theorem 3.3, we can obtain the errors of each component of the coefficient matrices A , B simultaneously, and assess the statistical significance of each predictor on each response. Using the ad-hoc approach of Yee and Hastie [6], a predictor can only be found to be significant across all responses. For example, Yee and Hastie [6] find the predictor `hurt` to be significant for all three groups (single, divorced/separated, widower). On the other hand, we can assess the significance of any response/predictor combination. Thus, we find `hurt` to be significant for single and divorced/separated groups, but not for widower men group (see Table 3).

A potential future direction for our approach was brought to our attention by a referee. The computational cost for fitting a reduced rank multinomial logistic regression can be very high. Powers et al. [20] proposed replacing the rank restriction with a restriction on the nuclear norm which amounts to a convex relaxation of the reduced-rank multinomial regression problem. Our methodology can be adapted to obtain asymptotic inference for the regularized parameter estimates.

Acknowledgement

Mariela Sued acknowledges the support of the Abdus Salam International Center for Theoretical Physics (ICTP), where part of this research was carried out.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported in part by FWF (Austrian Science Fund, <https://www.fwf.ac.at/en/>) [P30690-N35], Simons Foundation [360066], SECTEI (Secretaría de Estado de Ciencia, Tecnología e Innovación de la provincia de Santa Fe, <http://fich.unl.edu.ar/pagina/sectei-santa-fe/582/>) [2010-072-14], UNL [500-040, 501-499, 500-062], CONICET (El Consejo Nacional de Investigaciones Científicas y Técnicas, <http://www.conicet.gov.ar/?lan=en>) [PIP 742], ANPCYT (<http://www.agencia.mincyt.gob.ar/>) [PICT 2012-2590], Universidad de Buenos Aires [20020150100122ba].

Supplemental data and underlying research materials

The R script to reproduce the data analysis in Section 4 can be accessed at supplemental data.

References

- [1] Reinsel GC, Velu RP. Multivariate reduced rank regression. New York: Springer; 1998 (Lecture Notes in Statistics).
- [2] Anderson TW. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann Math Statist.* 1951;22:327–351.
- [3] Izenman AJ. Reduced-rank regression for the multivariate linear model. *J Multivariate Anal.* 1975;5:248–264.
- [4] Izenman AJ. Modern multivariate. Statistical techniques: regression, classification and manifold learning. New York: Springer; 2008.

[5] Fan J, Gong W, Zhu Z. Generalized high-dimensional trace regression via nuclear norm regularization; 2017 [cited 2017 Oct 23]. arXiv pre-print, available from: <https://arxiv.org/pdf/1710.08083.pdf>.

[6] Yee TW, Hastie TJ. Reduced-rank vector generalized linear models. *Stat Model*. 2003;3:15–41.

[7] Yee TW. Vector generalized linear and additive models. New York: Springer; 2015.

[8] Bura E, Duarte S, Forzani L. Sufficient reductions in regressions with exponential family inverse predictors. *J Amer Statist Assoc*. 2016;111:1313–1329.

[9] Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, No. 1; 1967. p. 221–233.

[10] Haberman SJ. Concavity and estimation. *Ann Statist*. 1989;17:1631–1661.

[11] Niemiro W. Asymptotics for M-estimators defined by convex minimization. *Ann Statist*. 1992;20:1514–1533.

[12] Hjort NL, Pollard D. Asymptotics for minimisers of convex processes; 2011. arXiv:1107.3806.

[13] Geyer CJ. On the asymptotics of constrained M-estimation. *Ann Statist*. 1994;22:1993–2010.

[14] van der Vaart AW. *Asymptotic statistics*. Cambridge: Cambridge University Press; 2000.

[15] Yee TW. VGAM: vector generalized linear and additive models. R package version 1.0-3. 2015 [cited 2018 Feb 7]. Available from: <https://CRAN.R-project.org/package=VGAM>.

[16] Wasserman L. *All of statistics: a concise course in statistical inference*. New York: Springer-Verlag; 2013.

[17] Duarte SL. *Modelos lineales generalizados: regresión de rango reducido y reducción suficiente de dimensiones [Tesis Doctoral]*. Argentina: Universidad Nacional del Litoral; 2016.

[18] Bura E, Yang J. Dimension estimation in sufficient dimension reduction: a unifying approach. *J Multivariate Anal*. 2011;102:130–142.

[19] Yee TW, Wild CJ. Vector generalized additive models. *J R Stat Soc Ser B*. 1996;58(3):481–493.

[20] Powers S, Hastie T, Tibshirani R. Nuclear penalized multinomial regression with an application to predicting at-bat outcomes in baseball. To appear in special edition ‘Statistical Modelling for Sports Analytics,’ *Statistical Modelling*; 2018.

[21] Cook RD, Ni L. Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *J Amer Statist Assoc*. 2005;100:410–428.

[22] Puntanen S, Styan GP, Isotalo J. *Matrix tricks for linear statistical models: our personal top twenty*. Berlin Heidelberg: Springer Science and Business Media; 2011.

Appendix. Proofs

The consistency of M-estimators has long been established (see, for instance, Theorem 5.7 in [14]). The functions $M(\xi)$ and $M_n(\xi)$ are defined in Equation (2). Typically, the proof for the consistency of M-estimators assumes that ξ_0 , the parameter of interest, is a *well-separated* point of maximum of M , which is ascertained by assumptions (a) and (b) of Proposition A.1. Assumption (c) of Proposition A.1 yields uniform consistency of M_n as an estimator of M , a property needed in order to establish the consistency of M-estimators.

Proposition A.1: *Assume (a) $m_\xi(z)$ is a strictly concave function in $\xi \in \Xi$, where Ξ is a convex open subset of a Euclidean space; (b) the function $M(\xi)$ is well defined for any $\xi \in \Xi$ and has a unique maximizer ξ_0 ; that is, $M(\xi_0) > M(\xi)$ for any $\xi \neq \xi_0$; and (c) for any compact set \mathcal{K} in Ξ ,*

$$E \left[\sup_{\xi \in \mathcal{K}} |m_\xi(Z)| \right] < \infty. \tag{A1}$$

Then, for each $n \in \mathbb{N}$, there exists a unique M-estimator $\hat{\xi}_n$ for the criterion function M_n over Ξ . Moreover, $\hat{\xi}_n \rightarrow \xi_0$ a.e. as $n \rightarrow \infty$.

Proof: For each compact subset \mathcal{K} of Ξ , $\{m_\xi : \xi \in \mathcal{K}\}$ is a collection of measurable functions which, by assumption (c), has an integrable envelope. Moreover, for each fixed z , the map $\xi \mapsto m_\xi(z)$ is continuous, since it is concave and defined on the open set Ξ . As stated in Example 19.8 of van der Vaart [14], these conditions guarantee that the class is Glivenko–Cantelli. That is,

$$\text{pr} \left(\lim_{n \rightarrow \infty} \sup_{\xi \in \mathcal{K}} |P_n m_\xi - P m_\xi| = 0 \right) = 1. \tag{A2}$$

We need to prove that there exists a unique maximizer of $M_n(\xi) = P_n \xi$, and that it converges to the maximizer of $M(\xi) = P m_\xi$. We first consider the deterministic case ignoring for the moment that $\{M_n\}$ is a sequence of random functions.

Since ξ_0 belongs to the open set Ξ , there exists $\varepsilon_0 > 0$ such that the closed ball $B[\xi_0, \varepsilon_0]$ is contained in Ξ . Uniform convergence of $\{M_n\}$ to M over $\mathcal{K} = \{\xi : \|\xi - \xi_0\| = \varepsilon_0\}$ guarantees that

$$\lim_{n \rightarrow \infty} \sup_{\|\xi - \xi_0\| = \varepsilon_0} \{M_n(\xi) - M_n(\xi_0)\} = \sup_{\|\xi - \xi_0\| = \varepsilon_0} \{M(\xi) - M(\xi_0)\} < 0 \tag{A3}$$

because $M(\xi_0) > M(\xi)$ for any $\xi \neq \xi_0$. Then, for $n \geq n_0(\varepsilon_0)$,

$$\sup_{\|\xi - \xi_0\| = \varepsilon_0} M_n(\xi) - M_n(\xi_0) < 0. \tag{A4}$$

Let $\zeta_n(\xi) = M_n(\xi) - M_n(\xi_0)$. Since M_n is concave and continuous, ζ_n attains its maximum over the compact set $B[\xi_0, \varepsilon_0]$, which we denote by $\hat{\xi}_n$. Note that $\zeta_n(\xi_0) = 0$ and ζ_n is strictly smaller than zero in the boundary of the ball, as shown in Equation (A4); therefore, we conclude that $\hat{\xi}_n \in B(\xi_0, \varepsilon_0)$, so that $\hat{\xi}_n$ is a local maximum for ζ_n .

Let ξ satisfy $\|\xi - \xi_0\| > \varepsilon_0$. The convexity of Ξ implies there exists $t \in (0, 1)$ such that $\tilde{\xi} = (1 - t)\xi_0 + t\xi$ satisfies $\|\tilde{\xi} - \xi_0\| = \varepsilon_0$, and therefore

$$\zeta_n(\hat{\xi}_n) \geq \zeta_n(\xi_0) = 0 > \zeta_n(\tilde{\xi}) \geq t\zeta_n(\xi) + (1 - t)\zeta_n(\xi_0) = t\zeta_n(\xi), \tag{A5}$$

implying that $\zeta_n(\xi) < 0 < \zeta_n(\hat{\xi}_n)$. Therefore, the maximum $\hat{\xi}_n \in B(\xi_0, \varepsilon_0)$ is global. The strict concavity of M_n guarantees that such global maximum is unique, thus $\hat{\xi}_n$ is the unique solution to the optimization problem in Equation (3). By repeating this argument for any $\varepsilon < \varepsilon_0$, we prove the convergence of the sequence $\{\hat{\xi}_n\}$ to ξ_0 .

Turning to the stochastic case, the uniform convergence of M_n to M over $\mathcal{K} = B[\xi_0, \varepsilon_0]$ on a set Ω_1 , with $\text{pr}(\Omega_1) = 1$, as assumed in Equation (A2), guarantees the deterministic result can be applied to any element of Ω_1 , which obtains the result. ■

Proof of Lemma 2.3: Let $\{\hat{\xi}_n\}$ be any (weak/strong) M-estimator sequence of the unconstrained maximization problem. Since $M_n(\hat{\xi}_n) \geq \sup_{\xi \in \Xi} M_n(\xi) - A_n$ with $A_n \rightarrow 0$, we have

$$1 = \lim_{n \rightarrow \infty} \text{pr} \left(\sup_{\xi \in \Xi} P_n m_\xi < \infty \right) \leq \lim_{n \rightarrow \infty} \text{pr} \left(\sup_{\xi \in \Xi^{\text{res}}} P_n m_\xi < \infty \right). \tag{A6}$$

Define

$$\Omega_n := \left\{ \sup_{\xi \in \Xi^{\text{res}}} P_n m_\xi < \infty \right\}. \tag{A7}$$

For all n , there exists $\hat{\xi}_n^{\text{res}}$ such that

$$M_n(\hat{\xi}_n^{\text{res}}) \geq \sup_{\xi \in \Xi^{\text{res}}} M_n(\xi) - \frac{1}{n^2} \tag{A8}$$

on Ω_n . Let $\hat{\xi}_n^{\text{res}} \doteq 0$ on Ω_n^c . Then, since $\text{pr}(\Omega_n) \rightarrow 1$, $\{\hat{\xi}_n^{\text{res}}\}$ is a strong M-estimator for the criterion function M_n over Ξ^{res} . ■

Proof of Proposition 2.4: In Proposition A.1 we established the existence of a unique maximizer $\hat{\xi}_n$ for the criterion function M_n over Ξ . We can now invoke Lemma 2.3 to guarantee the existence of $\hat{\xi}_n^{\text{res}}$, a strong M-estimator for the criterion function M_n over Ξ^{res} . Let $\{\hat{\xi}_n^{\text{res}}\}$ be any strong M-estimator for the criterion function M_n over Ξ^{res} . We start from the deterministic case:

$$M_n(\hat{\xi}_n^{\text{res}}) \geq \sup_{\xi \in \Xi^{\text{res}}} M_n(\xi) - A_n, \tag{A9}$$

where M_n is defined in Equation (2) and A_n is a sequence of real numbers with $A_n \rightarrow 0$. As in the proof of Proposition A.1, define $\zeta_n(\xi) = M_n(\xi) - M_n(\xi_0)$ to obtain that, for ε_0 small enough,

$$\sup_{\|\xi - \xi_0\| = \varepsilon_0} \zeta_n(\xi) \leq \frac{1}{2} \sup_{\|\xi - \xi_0\| = \varepsilon_0} \{M(\xi) - M(\xi_0)\} := -\delta(\varepsilon_0) \tag{A10}$$

for n large enough. Under Condition 2.2, $\xi_0 \in \Xi^{\text{res}}$, and therefore, by Equation (A9),

$$\zeta_n(\hat{\xi}_n^{\text{res}}) = M_n(\hat{\xi}_n^{\text{res}}) - M_n(\xi_0) \geq M_n(\hat{\xi}_n^{\text{res}}) - \sup_{\xi \in \Xi^{\text{res}}} M_n(\xi) \geq -A_n. \tag{A11}$$

Since $A_n \rightarrow 0$, $-A_n > -\delta(\varepsilon_0)$ for n large enough. Combining this with Equations (A10) and (A11) obtains

$$\sup_{\|\xi - \xi_0\| = \varepsilon_0} \zeta_n(\xi) < \zeta_n(\hat{\xi}_n^{\text{res}}).$$

We will deduce that $\|\hat{\xi}_n^{\text{res}} - \xi_0\| < \varepsilon_0$, once we prove that

$$\sup_{\|\xi - \xi_0\| = \varepsilon_0} \zeta_n(\xi) = \sup_{\|\xi - \xi_0\| \geq \varepsilon_0} \zeta_n(\xi). \tag{A12}$$

Now, let us prove Equation (A12). Choose ξ with $\|\xi - \xi_0\| > \varepsilon_0$, and take $t \in (0, 1)$ such that $\tilde{\xi} = (1 - t)\hat{\xi}_n + t\xi$ is a distance ε_0 from ξ_0 , where $\hat{\xi}_n$ is the maximizer of ζ_n over Ξ , as defined in Proposition A.1, which is assumed to be at distance smaller than ε_0 from ξ_0 . Then,

$$\zeta_n(\tilde{\xi}) = \zeta_n\left((1 - t)\hat{\xi}_n + t\xi\right) \geq (1 - t)\zeta_n(\hat{\xi}_n) + t\zeta_n(\xi) \geq \zeta_n(\xi).$$

Thus,

$$\sup_{\|\xi - \xi_0\| = \varepsilon_0} \zeta_n(\xi) \geq \zeta_n(\tilde{\xi}) \geq \zeta_n(\xi), \quad \text{for any } \xi \text{ with } \|\xi - \xi_0\| > \varepsilon_0,$$

which in turn yields (A12). When $A_n = o_p(1)$, convergence in probability of $\{\hat{\xi}_n^{\text{res}}\}$ to ξ_0 is equivalent to the existence of an almost everywhere convergent sub-sub-sequence for any subsequence $\{\hat{\xi}_{n_k}^{\text{res}}\}$. Therefore, by applying the deterministic result to the set of probability one, where there exists a sub-subsequence of A_{n_k} that converges to zero a.e. we obtain the result. ■

Regularity conditions for Proposition 2.7 (from Theorem 5.23, p. 53 in [14]).

Condition A.2: For each ξ in an open subset Ξ of a Euclidean space, $m_\xi(z)$ is a measurable function in z such that $m_\xi(z)$ is differentiable in ξ_0 for almost every z with derivative $\dot{m}_{\xi_0}(z)$.

Condition A.3: There exists a measurable function $\phi(z)$ with $P\phi^2 < \infty$ such that, for any ξ_1 and ξ_2 in a neighbourhood of ξ_0 ,

$$|m_{\xi_1}(z) - m_{\xi_2}(z)| \leq \phi(z)\|\xi_1 - \xi_2\|. \tag{A13}$$

Condition A.4: The map $\xi \rightarrow Pm_\xi$ admits a second-order Taylor expansion at a point of maximum ξ_0 with non-singular symmetric second derivative matrix V_{ξ_0} .

Under regularity conditions A.2–A.4, van der Vaart proved in Theorem 5.23 of his book [14] that if $\{\hat{\xi}_n\}$ is a strong M-estimator sequence for the criterion function $P_n m_\xi$ over Ξ and $\hat{\xi}_n \rightarrow \xi_0$ in probability, then

$$\sqrt{n}(\hat{\xi}_n - \xi_0) = -V_{\xi_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{m}_{\xi_0}(Z_i) + o_p(1). \tag{A14}$$

Moreover, $\sqrt{n}(\hat{\xi}_n - \xi_0)$ is asymptotically normal with mean zero and

$$\text{avar} \left\{ \sqrt{n}(\hat{\xi}_n - \xi_0) \right\} = V_{\xi_0}^{-1} P\dot{m}_{\xi_0} \dot{m}_{\xi_0}^T V_{\xi_0}^{-1}. \tag{A15}$$

This result will be invoked in the following proofs.

Proof of Proposition 2.7: Assume that $\{\hat{\xi}_n^{\text{res}}\}$ is a sequence in Ξ^{res} that converges in probability to $\xi_0 \in \mathcal{M}$, which is assumed to be open in Ξ^{res} . Then, $\text{pr}(\hat{\xi}_n^{\text{res}} \in \mathcal{M}) \rightarrow 1$. Bicontinuity of h guarantees that $s_n^* = h^{-1}(\hat{\xi}_n^{\text{res}})$ converges in probability to $s_0 = h^{-1}(\xi_0)$. Note that

$$P_n m_{h(s_n^*)} = P_n m_{\hat{\xi}_n^{\text{res}}} \geq \sup_{\xi \in \Xi^{\text{res}}} P_n m_\xi \geq \sup_{\xi \in \mathcal{M}} P_n m_\xi \geq \sup_{s \in \mathcal{S}} P_n m_{h(s)}, \tag{A16}$$

except for an $o_p(n^{-1})$ term that is omitted in the last three inequalities. Therefore, $\{s_n^*\}$ is a strong maximizing sequence for the criterion function $P_n m_{h(s)}(z)$ over \mathcal{S} .

We next verify Conditions A.2–A.4 are satisfied for $\{s_n^*\}$, s_0 , $m_{h(s)}(z)$ and $Pm_{h(s)}$. Specifically, Condition A.2 holds since $m_{h(s)}$ is a measurable function in z for all $s \in \mathcal{S}$ and $m_{h(s)}(z)$ is differentiable in s_0 for almost every z . In fact, $h(s_0) = \xi_0$, $m_\xi(z)$ is differentiable at ξ_0 and $h(s)$ is also differentiable. Moreover, the derivative function is $\nabla h(s_0)\dot{m}_{\xi_0}$.

For all s_1 and s_2 in a neighbourhood of s_0 , by the continuity of h , $h(s_1)$ and $h(s_2)$ are in a neighbourhood of ξ_0 . Then

$$\begin{aligned} |m_{h(s_1)}(z) - m_{h(s_2)}(z)| &\leq \phi(z)\|h(s_1) - h(s_2)\| \\ &\leq \phi(z)\|\nabla h\|_{\infty, \mathcal{N}_{s_0}} \|s_1 - s_2\|, \end{aligned}$$

where $\|\nabla h\|_{\infty, \mathcal{N}_{s_0}}$ denotes the maximum of $\|\nabla h(s)\|$ in a neighbourhood \mathcal{N}_{s_0} of s_0 . The first inequality holds because such condition is valid in the unconstrained problem and the second inequality follows since h is continuously differentiable at s_0 . Thus, the Lipschitz Condition A.3 is satisfied.

For Condition A.4, we observe that the function $s \mapsto Pm_{h(s)}$ is twice continuously differentiable in s_0 because both Pm_ξ and $h(s)$ satisfy the required regularity properties at ξ_0 and s_0 , respectively. Moreover, since $Pm_{\xi_0} = 0$, the second derivative matrix of $Pm_{h(s)}$ at s_0 , is $W_{s_0} = \nabla h(s_0)^T V_{\xi_0} \nabla h(s_0)$, where V_{ξ_0} is the second derivative matrix of Pm_ξ at ξ_0 . The matrix W_{s_0} is non-singular and symmetric because $\nabla h(s_0)$ is full rank and V_{ξ_0} is non-singular and symmetric.

We can now apply Theorem 5.23 in [14], and obtain

$$\sqrt{n}(s_n^* - s_0) = - \left(\nabla h(s_0)^T V_{\xi_0} \nabla h(s_0) \right)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla h(s_0)^T \dot{m}_{\xi_0}(Z_i) + o_p(1), \tag{A17}$$

so that the first-order Taylor series expansion of $h(s_n^*)$ around s_0 is

$$\begin{aligned} \sqrt{n}(h(s_n^*) - h(s_0)) &= \nabla h(s_0) \sqrt{n}(s_n^* - s_0) + o_p(1) \\ &= - \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla h(s_0) \left(\nabla h(s_0)^T V_{\xi_0} \nabla h(s_0) \right)^{-1} \nabla h(s_0)^T \dot{m}_{\xi_0}(Z_i) + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \Pi_{\xi_0(-V_{\xi_0})} IF_{\xi_0}(Z_i) + o_p(1), \end{aligned}$$

for $\Pi_{\xi_0(-V_{\xi_0})}$ and $IF_{\xi_0}(z)$ defined in Equations (6) and (5), respectively, which obtains the expansion in Equation (7). Now, Equation (8) follows immediately. ■

Proof of Theorem 3.1: Write

$$\eta_{x1} = \bar{\eta}_1 + \beta x = (\bar{\eta}_1, \beta) f(x) = (f(x)^T \otimes I_{k_1}) \text{vec}(\bar{\eta}_1, \beta),$$

where $f(x)^T = (1, x^T) \in \mathbb{R}^{1 \times (\rho+1)}$. Then, in matrix form,

$$\eta_x = \begin{pmatrix} \bar{\eta}_1 + \beta x \\ \bar{\eta}_2 \end{pmatrix} = \begin{pmatrix} f(x)^T \otimes I_{k_1} & 0 \\ 0 & I_{k_2} \end{pmatrix} \begin{pmatrix} \text{vec}(\bar{\eta}_1, \beta) \\ \bar{\eta}_2 \end{pmatrix} = F(x)\xi, \tag{A18}$$

where

$$F(x) = \begin{pmatrix} f(x)^T \otimes I_{k_1} & 0 \\ 0 & I_{k_2} \end{pmatrix} \in \mathbb{R}^{k \times (k_1(\rho+1) + k_2)},$$

and $\xi = (\bar{\eta}_1^T, \text{vec}^T(\beta), \bar{\eta}_2^T)^T$ is the vector of parameters of model (13). Note that $F(x)\xi \in H$ for any value of ξ with H defined in Equation (10). This notation allows to simplify the expression for the log-likelihood function in Equation (15), and replace it with

$$m_\xi(z) = T(y)^T F(x)\xi - \psi(F(x)\xi). \tag{A19}$$

The regularity conditions required to derive consistency and asymptotic distribution of the MLE are: For any ξ and any $\bar{\eta}$ and for any compact $\mathcal{K} \subset \Xi = \mathbb{R}^{k_1} \times \mathbb{R}^{k_1 \rho} \times H_2$.

$$\text{pr}(F(X)\xi = \bar{\eta}) < 1. \tag{A20}$$

$$\text{E} \left[\|T(Y)^T F(X)\| \right] < \infty, \quad \text{E} \left[\sup_{\xi \in \mathcal{K}} |\psi(F(X)\xi)| \right] < \infty, \tag{A21}$$

$$\text{E} \left[\|T(Y)^T F(X)\|^2 \right] < \infty, \quad \text{E} \left[\sup_{\xi \in \mathcal{K}} \|\nabla \psi(F(X)\xi) F(X)\|^2 \right] < \infty, \tag{A22}$$

$$\text{E} \left[\sup_{\xi \in \mathcal{K}} \|F(X)^T \nabla^2 \psi(F(X)\xi) F(X)\| \right] < \infty. \tag{A23}$$

To prove the existence, uniqueness and consistency of the MLE under the present model, $\hat{\xi}_n$, we next show that the assumptions stated in Proposition A.1 are satisfied.

The strict convexity of ψ implies that, for each fixed z , $m_\xi(z)$ is a strictly concave function in $\xi \in \Xi = \mathbb{R}^{k_1 + \rho k_1} \times H_2$. The concavity of $m_\xi(z)$ is preserved under expectation, thus $M(\xi) = Pm_\xi$ is concave. The identifiability condition

satisfied by the exponential family in Section 3.1 allows applying Lemma 5.35 of van der Vaart [14, p.62] to conclude that

$$E \left[T(Y)^T F(Y)\xi - \psi(F(X)\xi) \mid X \right] \leq E \left[T(Y)^T F(X)\xi_0 - \psi(F(X)\xi_0) \mid X \right]. \tag{A24}$$

Taking expectation with respect to X , we conclude that $Pm_\xi \leq Pm_{\xi_0}$ for any ξ . Moreover, if $Pm_{\xi_1} = Pm_{\xi_0}$, $\text{pr}(F(X)\xi_1 = F(X)\xi_0) = 1$, which contradicts the hypothesis (A20). Finally, the integrability of Equation (A1) follows from Equation (A21).

The conditions A.2–A.4 required by van der Vaart’s Theorem 5.23 to derive the asymptotic distribution of M-estimators are easily verifiable under the integrability assumptions stated in Equations (A22) and (A23).

The second derivative matrix of Pm_ξ at ξ_0 is

$$V_{\xi_0} = -E \left[F(X)^T \nabla^2 \psi(F(X)\xi_0) F(X) \right]. \tag{A25}$$

Finally, observe that

$$\begin{aligned} P \left[\hat{m}_{\xi_0} \hat{m}_{\xi_0}^T \right] &= E \left[F(X)^T \left\{ T(Y) - \nabla^T \psi(F(X)\xi_0) \right\} \left\{ T(Y)^T - \nabla \psi(F(X)\xi_0) \right\} F(X) \right] \\ &= E \left\{ F(X)^T \nabla^2 \psi(F(X)\xi_0) F(X) \right\} \end{aligned}$$

to deduce that, according to the general formula for the asymptotic variance of an M-estimator in (A15), the asymptotic variance of $\sqrt{n}(\hat{\xi}_n - \xi_0)$ is given by (16). ■

Lemmas A.5–A.10 and Corollary A.6 are required to prove Proposition 3.2.

Lemma A.5: Assume that $\beta_{01} \in \mathbb{R}_{\text{first},d}^{k_1 \times r}$ and can be written as in Equation (21). Let $(S_0, T_0) \in \mathbb{R}_d^{k_1 \times d} \times \mathbb{R}_d^{d \times r}$ with $S_0 T_0 = \beta_{01}$. Then, there exists $U \in \mathbb{R}_d^{d \times d}$ so that $S_0 = S_0(U)$ and $T_0 = T_0(U)$, with

$$S_0(U) := \begin{pmatrix} U \\ A_0 U \end{pmatrix} \quad \text{and} \quad T_0(U) := U^{-1} B_0. \tag{A26}$$

Proof: Let

$$S_0 = \begin{pmatrix} S_{01} \\ S_{02} \end{pmatrix},$$

with $S_{01} \in \mathbb{R}^{d \times d}$ and $S_{02} \in \mathbb{R}^{(k_1-d) \times d}$. The matrix $S_{01} T_0$ is comprised of the first d rows of β_{01} which are linearly independent. Then, $d = \text{rank}(S_{01} T_0) \leq \text{rank}(S_{01}) \leq d$, hence S_{01} is invertible. Take $U = S_{01}$. From the expression (21) for β_{01} , we have

$$\begin{aligned} S_{01} T_0 &= B_0 \\ S_{02} T_0 &= A_0 B_0. \end{aligned}$$

Thus, $T_0 = U^{-1} B_0$, and since $T_0 T_0^T \in \mathbb{R}^{d \times d}$ and $\text{rank } d$,

$$S_{02} = A_0 B_0 T_0^T (T_0 T_0^T)^{-1} = A_0 B_0 B_0^T U^{-1} (U^{-1} B_0 B_0^T U^{-1})^{-1} = A_0 U. \tag{A27}$$

Corollary A.6: Let $\beta_{01} \in \mathbb{R}_{\text{first},d}^{k_1 \times r}$ and can be written as in Equation (21). For U in $\mathbb{R}_d^{d \times d}$ and $S_0(U), T_0(U)$ as defined in Equation (A26), the pre-image of ξ_0 through $g, g^{-1}(\xi_0)$, satisfies

$$g^{-1}(\xi_0) \cong \left\{ (\bar{\eta}_{01}, S_0(U), T_0(U), \beta_{02}, \bar{\eta}_{02}) : U \in \mathbb{R}_d^{d \times d} \right\}. \tag{A27}$$

Lemma A.7: $\mathbb{R}_d^{d \times m}$ with $d \leq m$ is an open set in $\mathbb{R}^{d \times m}$.

Proof: We will show that the complement of $\mathbb{R}_d^{d \times m}$ is closed. Consider $(T_n)_{n \geq 1} \in \mathbb{R}^{d \times m}$, each T_n of rank strictly less than d , and assume that T_n converges to $T \in \mathbb{R}^{d \times m}$. Note that, $|T_n T_n^T| = 0$ for all $n \geq 1$, so that $|T T^T|$ is also equal to zero. Hence, $\text{rank}(T) < d$. ■

Lemma A.8: Θ and S are open sets.

Proof: Up to a homeomorphism, Θ and \mathcal{S} are equivalent to

$$\mathbb{R}^{k_1} \times \mathbb{R}_d^{k_1 \times d} \times \mathbb{R}_d^{d \times r} \times \mathbb{R}^{k_1 \times (p-r)} \times H_2 \quad \text{and} \quad \mathbb{R}^{k_1} \times \mathbb{R}^{(k_1-d) \times d} \times \mathbb{R}_d^{d \times r} \times \mathbb{R}^{k_1 \times (p-r)} \times H_2,$$

respectively, which are products of open sets by Lemma A.7. ■

Lemma A.9: $h : \mathcal{S} \rightarrow \mathcal{M}$ is one to one bicontinuous.

Proof: It suffices to note that $h_1 : \mathbb{R}^{(k_1-d) \times d} \times \mathbb{R}_d^{d \times r} \mapsto \mathbb{R}_{\text{first},d}^{k_1 \times r}$ with

$$(A, B) \rightarrow \begin{pmatrix} B \\ AB \end{pmatrix}$$

satisfies the required properties for h . Let $h_1^{-1} : \mathbb{R}_{\text{first},d}^{k_1 \times r} \mapsto \mathbb{R}^{(k_1-d) \times d} \times \mathbb{R}_d^{d \times r}$ with

$$\begin{pmatrix} B \\ C \end{pmatrix} \rightarrow (CB^T (BB^T)^{-1}, B). \tag{A28}$$

Then, $h_1^{-1}(h_1(A, B)) = (A, B)$ and

$$h_1 \left(h_1^{-1} \begin{pmatrix} B \\ AB \end{pmatrix} \right) = \begin{pmatrix} B \\ AB \end{pmatrix},$$

and, therefore, h_1^{-1} is the inverse of h_1 . Thus, h_1 is one to one and bicontinuous. ■

Lemma A.10: \mathcal{M} is open in Ξ^{res}

Proof: It suffices to show $\mathbb{R}_{\text{first},d}^{k_1 \times r}$ is an open set in $\mathbb{R}_d^{k_1 \times r}$, or equivalently, that the complement of $\mathbb{R}_{\text{first},d}^{k_1 \times r}$ is a closed set in $\mathbb{R}_d^{k_1 \times r}$. Let $\{T_n\} \subset \mathbb{R}_d^{k_1 \times r}$ be a sequence such that the first d rows are not linearly independent and T_n converges to $T \in \mathbb{R}_d^{k_1 \times r}$. Then, if we write

$$T_n = \begin{pmatrix} T_{n_1} \\ T_{n_2} \end{pmatrix},$$

with $T_{n_1} \in \mathbb{R}^{d \times r}$ and $T_{n_2} \in \mathbb{R}^{(k_1-d) \times r}$, we obtain that $|T_{n_1} T_{n_1}^T| = 0$ for all n . Then $|T_1 T_1^T| = 0$, where T_1 comprises of the first d rows of T , and therefore the first d rows of T are not linearly independent. ■

Proof of Proposition 3.2: We verify that Conditions 2.2, 2.5 and 2.6 are satisfied.

Condition 2.2: By Lemma A.8, the set Θ , defined in Equation (20), is open and ξ_0 belongs to $g(\Theta) = \Xi^{\text{res}}$.

Condition 2.5: Since the first d rows of β_{01} are linearly independent, $\xi_0 \in \mathcal{M}$. Then, applying Lemma A.10 obtains that \mathcal{M} is an open set in Ξ^{res} .

Consider (\mathcal{S}, h) as in Equation (23). By Lemma A.8, \mathcal{S} is an open set, and h is one-to-one and bicontinuous by Lemma A.9. Furthermore, if

$$s_0 = (\bar{\eta}_{01}, A_0, B_0, \beta_{02}, \bar{\eta}_{02}) \in \mathcal{S},$$

where A_0 y B_0 are given in Equation (21), then $h(s_0) = (\bar{\eta}_{01}, \beta_{01}, \beta_{02}, \bar{\eta}_{02}) = \xi_0$.

The function h is twice continuously differentiable and its Jacobian is full rank. In fact, the latter is

$$\nabla h(\bar{\eta}_1, A, B, \beta_2, \bar{\eta}_2) = \begin{pmatrix} I_{k_1} & 0 & 0 & 0 & 0 \\ 0 & B^T \otimes \begin{pmatrix} 0 \\ I_{k_1-d} \end{pmatrix} & I_r \otimes \begin{pmatrix} I_d \\ A \end{pmatrix} & 0 & 0 \\ 0 & 0 & 0 & I_{k_1(p-r)} & 0 \\ 0 & 0 & 0 & 0 & I_{k_2} \end{pmatrix} \tag{A29}$$

of order $(k_1 p + k) \times (d(k_1 + r - d) + k_1(p - r) + k)$ with full column rank $d(k_1 + r - d) + k_1(p - r) + k$ (see [21]; also it is a direct implication of Theorem 5 in [22]).

Condition 2.6: Consider $\theta_0 \in g^{-1}(\xi_0)$, associated with U , as shown in Lemma A.5. Since

$$\nabla g(\bar{\eta}_1, S, T, \beta_2, \bar{\eta}_2) = \begin{pmatrix} I_{k_1} & 0 & 0 & 0 & 0 \\ 0 & T^T \otimes I_{k_1} & I_r \otimes S & 0 & 0 \\ 0 & 0 & 0 & I_{k_1(p-r)} & 0 \\ 0 & 0 & 0 & 0 & I_{k_2} \end{pmatrix}$$

then,

$$\begin{aligned} &\nabla g\left(\bar{\eta}_{10}, \begin{pmatrix} U \\ A_0 U \end{pmatrix}, U^{-1}B_0, \beta_{20}, \bar{\eta}_{20}\right) \\ &= \begin{pmatrix} I_{k_1} & 0 & 0 & 0 & 0 \\ 0 & B_0^T \otimes I_{k_1} & I_r \otimes \begin{pmatrix} I_d \\ A_0 \end{pmatrix} & 0 & 0 \\ 0 & 0 & 0 & I_{k_1(p-r)} & 0 \\ 0 & 0 & 0 & 0 & I_{k_2} \end{pmatrix} G, \end{aligned}$$

where

$$G = \begin{pmatrix} I_{k_1} & 0 & 0 & 0 & 0 \\ 0 & U^{-T} \otimes I_{k_1} & 0 & 0 & 0 \\ 0 & 0 & (I_r \otimes U) & 0 & 0 \\ 0 & 0 & 0 & I_{k_1(p-r)} & 0 \\ 0 & 0 & 0 & 0 & I_{k_2} \end{pmatrix}. \tag{A30}$$

Since G is invertible of order $(d(k_1 + r) + k_1(p - r) + k) \times (d(k_1 + r) + k_1(p - r) + k)$,

$$\text{span} \nabla g(\theta_0) = \text{span} \begin{pmatrix} I_{k_1} & 0 & 0 & 0 & 0 \\ 0 & B_0^T \otimes I_{k_1} & I_r \otimes \begin{pmatrix} I_d \\ A_0 \end{pmatrix} & 0 & 0 \\ 0 & 0 & 0 & I_{k_1(p-r)} & 0 \\ 0 & 0 & 0 & 0 & I_{k_2} \end{pmatrix}. \tag{A31}$$

Next, since $\nabla h(s_0) = \nabla g(\theta_0)G^{-1}\tilde{I}$, with

$$\tilde{I} = \text{diag} \left(I_{k_1} \quad I_d \otimes \begin{pmatrix} 0 \\ I_{k_1-d} \end{pmatrix} \quad I_{rd} \quad I_{k_1(p-r)} \quad I_{k_2} \right),$$

we have $\text{span} \nabla h(s_0) \subset \text{span} \nabla g(\theta_0)$. Applying again Theorem 5 in [22] obtains $\text{rank}(\nabla g(\theta_0)) = d(k_1 + r - d) + k_1(p - r) + k = \text{rank}(\nabla h(s_0))$. Therefore, $\text{span} \nabla h(s_0) = \text{span} \nabla g(\theta_0)$. ■

Proof of Theorem 3.3: In Theorem 3.1 we verified that the conditions of Theorem 5.23 in [14] are satisfied for the maximum likelihood estimator under model (12) satisfying (13). The asymptotic variance of the MLE estimator is given in Equation (16). We also showed Conditions 2.2–2.6 are satisfied in the proof of Proposition 3.2. The result follows from Proposition 2.7 since $-V_{\xi_0} = W_{\xi_0}^{-1}$ and

$$\begin{aligned} \text{avar}\{\sqrt{n}(\hat{\xi}_n^{\text{res}} - \xi_0)\} &= \Pi_{\xi_0(W_{\xi_0}^{-1})} W_{\xi_0} \Pi_{\xi_0(W_{\xi_0}^{-1})}^T \\ &= \nabla g(\theta_0)(\nabla g(\theta_0)^T W_{\xi_0}^{-1} \nabla g(\theta_0))^\dagger \nabla g(\theta_0)^T \\ &= \Pi_{\xi_0(W_{\xi_0}^{-1})} W_{\xi_0}. \end{aligned}$$

Now, by Equation (21),

$$C_0 = \begin{pmatrix} I_d \\ A_0 \end{pmatrix},$$

$\text{span} \nabla g(\theta_0)$ in Equation (A31) is equal to $\text{span}(G)$, with G defined in Equation (26). The result follows if we prove that $\text{span}(G)$ does not depend on the decomposition of β_{01} .

Suppose that $\beta_{01} = C_1 B_1$. Then $C_1 = C_0 M$ and $B_1 = M^{-1} B_0$, for an invertible $M \in \mathbb{R}^{d \times d}$. Let G_1 be the matrix corresponding to G using the new decomposition of β_{01} . Then, $G_1 = GH$, with

$$H = \begin{pmatrix} I_{k_1} & 0 & 0 & 0 & 0 \\ 0 & M^{-T} \otimes I_{k_1} & I_r \otimes M & 0 & 0 \\ 0 & 0 & 0 & I_{k_1(p-r)} & 0 \\ 0 & 0 & 0 & 0 & I_{k_2} \end{pmatrix},$$

where H is invertible. Therefore, $\text{span}(G_1) = \text{span}(G)$ and the result follows. ■